# 4 Detailed Analyses of Individual Genes

The main indicators of positive selection used so far in this thesis, high frequency derived alleles and high levels of population differentiation in individual nonsense-SNPs, are indirect and can readily arise in other ways as well. This chapter describes data from resequencing two examples of interesting genes, *CASP12* and *MAGEE2*, so that additional tests could be used to investigate whether the unusual characteristics were found in extended regions of DNA surrounding each nonsense-SNP and if so were likely to have arisen by neutral processes, or whether positive selection would provide the best explanation.

The *MAGEE2* gene came up as an interesting outlier in our genome-wide survey of nonsense-SNPs described in chapter 3. The *CASP12* gene was, however, analysed before embarking on the main part of this study and the results were published by Xue et al (2006). The genotyping of nonsense-SNP rs497116 in *CASP12* and the subsequent analysis was performed by myself, while the resequencing part was performed by Yali Xue who then performed the analysis on the variation data.

In this chapter, I will refer to the different nonsense-SNP alleles in *CASP12* as "inactive" and "active" (instead of "stop" and "normal") as the functional consequences of the mutation is known. For the *MAGEE2*, for which I have no functional information, I will continue to refer to the stop and normal allele at the nonsense-SNP as was done in the previous chapter.

## 4.1 Results

### 4.1.1 CASP12

The human caspase-12 gene (*CASP12*) is on chromosome 11 and has been shown to modulate inflammation and innate immunity in humans (Saleh et al. 2004). The variation at the nonsense-SNP, rs497116, in the *CASP12* gene produces two versions of the protein which exist in human populations (see Figure 30), a truncated inactive

version and a full-length active version (Fischer et al. 2002). We find that the gene is truncated by ~64% and that the nonsense-SNP is expected to trigger NMD. A previous study (Saleh et al. 2004) found that the active ancestral version (considered by them an unusual 'long' variant) was only present in populations of African descent and was associated with a reduction in levels of cytokines after stimulation by bacterial lipopolysaccharides, leading to a lower initial immune response. However, carrying the active allele was found to increase susceptibility to developing severe sepsis, a later over-reaction of the immune system, as well as resulting in higher mortality rates once sepsis had developed. The truncated derived version, on the other hand, was associated with lower levels of severe sepsis and was found to be nearly fixed in human populations.

As there was a limited amount of information available on the evolutionary history of *CASP12*, we decided to investigate whether the inactive form had spread by neutral genetic drift or whether this was a case of selective advantage associated with gene loss. The *CASP12* nonsense-SNP did not show up as an extreme outlier in our survey of nonsense-SNP, except for a high DAF and moderate $F_{ST}$ value of 0.24, which was not found to be significant when compared to empirical distributions.

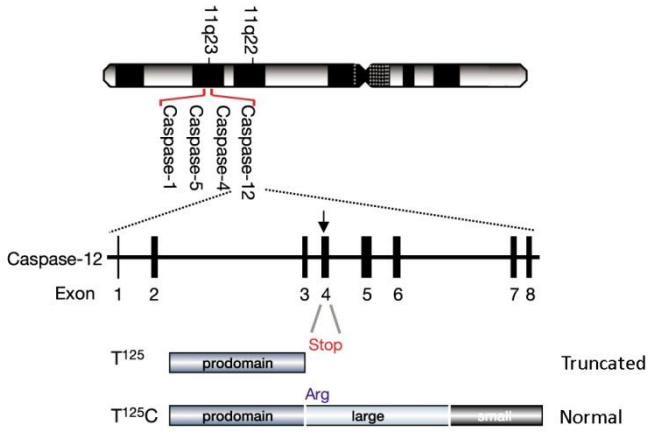


**Figure 30 Map of the region at 11q23 containing the *CASP12* gene.** The exon–intron organization of *CASP12* is shown. The arrow indicates the nonsense-SNP (rs497116) which changes an Argenine residue into a premature stop codon. The two products resulting, a truncated inactive version caused by the stop allele and a full-length active version, are displayed. This figure is adapted from (Saleh et al. 2004).

In accordance with previous results, we found the inactive allele to be nearly fixed in the human species with an overall frequency of 96%, while the active allele was mainly found in African populations (see Figure 31). Mbuti Pygmies and San have the highest frequencies of the active allele – 60% and 57%, respectively. Outside Africa, the active allele was very rare but was detected at low frequencies in Israel, Pakistan and China. No disagreement with HWE was observed in individual populations, but the pooled sample departed significantly from HWE (Chi-squared, P<<0.001), reflecting population subdivision.

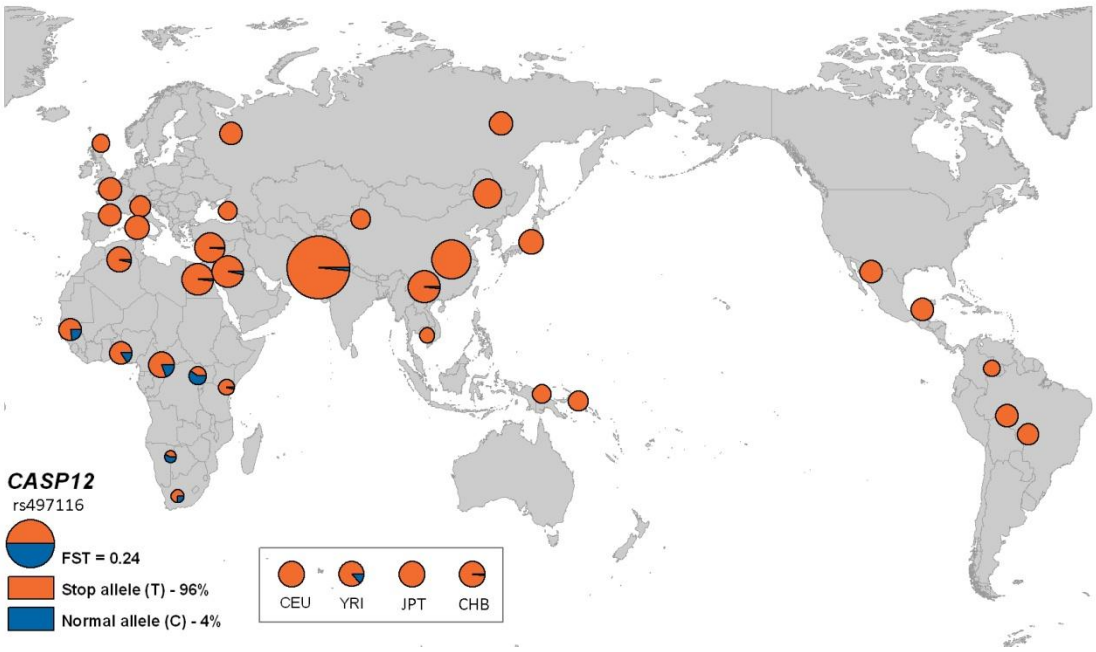


**Figure 31 Geographical distribution of inactive ("stop") and active ("normal") alleles of the CASP12 nonsense-SNP (rs rs497116).** The stop allele is nearly fixed in the human population and the normal allele is mainly found in African populations. The stop allele is represented in orange and the normal allele in blue. Pies are proportional to sample sizes.

#### *4.1.1.1 Sequence Variation in CASP12*

We next wanted to determine whether the observed predominance of the inactive allele was due to positive selection, or if it was the result of a neutral variant rising in frequency, for example because of the bottleneck associated with the human migration out of Africa. To this end, we resequenced a 13.3-kb region that covers the whole *CASP12* gene and an additional ~0.7 kb on each side of it in 77 individuals

from three HapMap populations (26 YRI, 26 CHB and 25 CEU). Our sample thus consisted of 155 chromosomes (154 from the HapMap samples in addition to the reference sequence which is of unknown origin). Eight were found to carry the active allele – six YRI, one CHB and the reference sequence, which is similar to the worldwide geographical distribution observed in Figure 31. The remaining 147 chromosomes carried the inactive allele. A total of 123 SNPs were detected (Appendix F.1.) but these were distributed very unevenly among the different versions of the gene and populations. We inferred the haplotypes from the SNP data and found that the active genes were much more diverse: the eight chromosomes carried 61 SNPs and showed a nucleotide diversity ($\pi$) of $19.7\times10^{-4}$, whereas the 147 inactive chromosomes carried 76 SNPs and had a nucleotide diversity almost 10 times lower, $2.0\times10^{-4}$ (see summary in Table 12). This led to higher diversity in the YRI ($\pi = 9.1\times10^{-4}$) than in the other populations, $1.9\times10^{-4}$ and $0.5\times10^{-4}$ in the CHB and CEU, respectively—a ratio more extreme than any encountered in a study of 132 genes in African American and European American populations (Akey et al. 2004). The inactive version was also more diverse in Africa than outside ($\pi = 4.4\times10^{-4}$ and $\pi = 0.7\times10^{-4}$) which is in accordance with most other studies on diversity within and outside Africa (Prugnolle et al. 2005; Rosenberg et al. 2002; The International HapMap Consortium 2005). The low diversity of the inactive form, particularly outside Africa, provided the second indication that their spread might have been rapid and thus due to positive selection.

#### *4.1.1.2 Long-Range Haplotype Tests (CASP12)*

We performed the REHH test (Sabeti et al. 2002) on four HapMap populations, CEU, CHB+JPT and YRI (described in section 2.3.8.4), and did not find any evidence for unusually extended haplotypes at high frequencies in *CASP12*. This test compares the suspected positively selected allele to the other allele at the same position and therefore relies on the SNP being polymorphic. As the inactive allele is fixed in CEU it is not possible to calculate REHH for this population. The inactive allele is also

nearly fixed in the Asian populations as only one active allele is reported in the combined populations of CHB and JPT. Again it is impossible to make inferences from such a low frequency. The YRI, however, have 14 copies of the active allele and thus it is possible to visualize the haplotypes in Haplotter (Figure 32) which is based on a LRH test which calculates the Integrated Haplotype Score (iHS) (Voight et al. 2006). In Figure 32A, a continuous block of the same colour represents a haplotype, and if it is shared by many chromosomes it will be thick. Indeed, there is some indication of such a block for the inactive allele (represented in red), but this is seen mainly on one side and does not include all chromosomes. However, while we see some indication of a long haplotype as was reported by Xue et al (2006), it must be compared with the ancestral allele haplotypes (blue). This also shows a long haplotype on one side for a proportion of chromosomes and thus LD seems to be similar for the active and inactive allele (Figure 32B).

We therefore conclude that LRH tests do not give us any evidence for positive selection of the *CASP12* nonsense-SNP. The low frequency of the active allele does not allow us to apply such tests in any population except the YRI, where it appears that either no such signature was formed, or enough time has passed for recombination to break up the long range structure. In fact, the age of the inactive allele was estimated between ~100-500 KYA in Xue *et al* (2006) which is too old for the LRH tests to detect a signal.

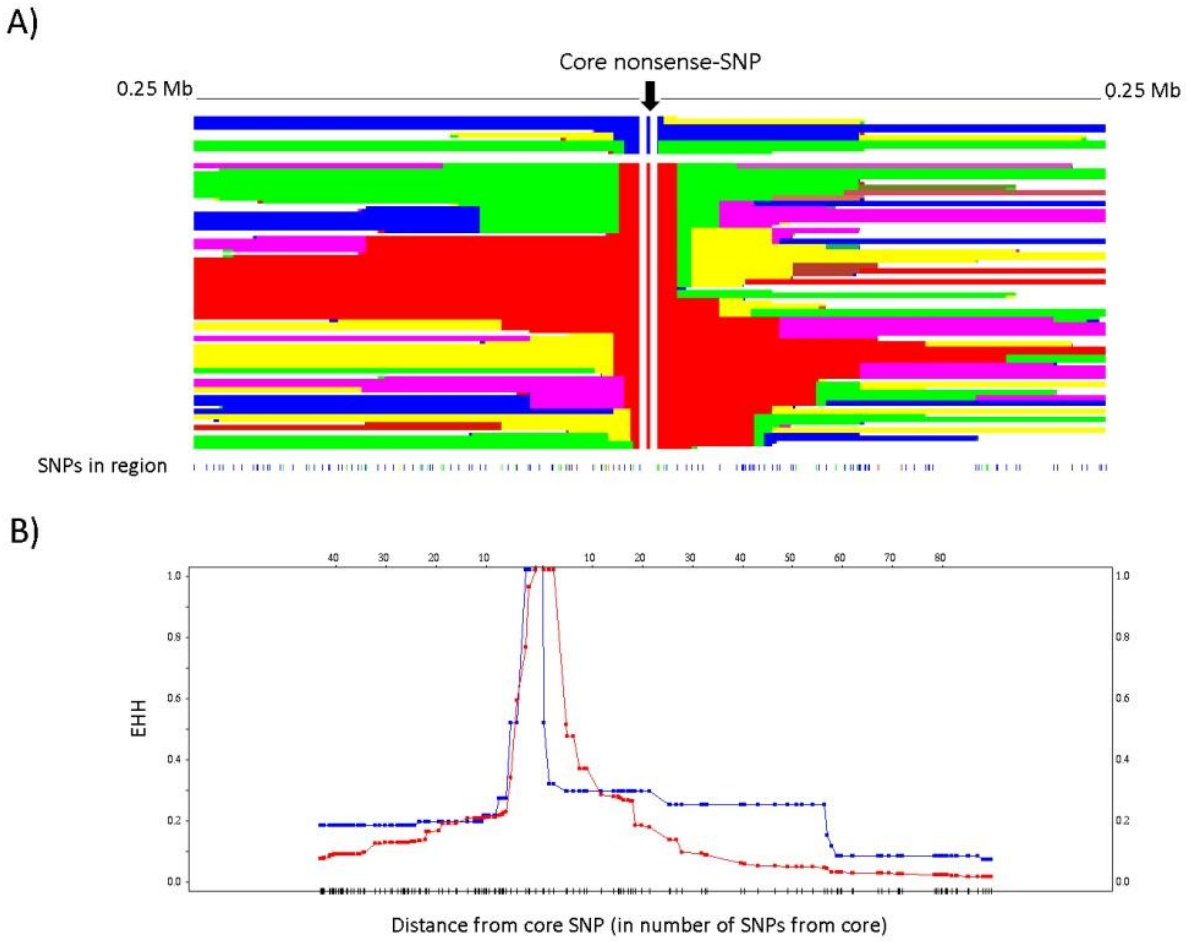


**Figure 32 *CASP12* haplotypes in YRI A) Haplotypes at different distances from the nonsense-SNP (core) at the centre.** Each horizontal line represents the haplotype of each chromosome. The blue vertical line represents the ancestral state (active allele) and the derived state (inactive allele) is represented in red. The distances over which the haplotypes are spread is displayed at the top of the graph. The total region size displayed on the top is 0.5 Mb and the SNPs in the region are showed at the bottom. **B) The decay of Extended Haplotype Homozygosity (EHH) at different distances from the nonsense-SNP (core).** The decay starts increasing at a short distance from the nonsense-SNP, for both the inactive (red) and active (blue) allele.

#### *4.1.1.3 Neutrality Tests (CASP12)*

Neutral models of evolution provide predictions of expected allele-frequency characteristics, and observed patterns can be compared with these. We calculated Tajima's *D* (Tajima 1989c), Fu and Li's *D* and *F* (Fu and Li 1993), and Fay and Wu's *H* (Fay and Wu 2000). The results are summarized in Table 12. Neutrality is rejected for *CASP12* by all tests in the combined populations. In individual populations, neutrality is similarly rejected by all tests for the CHB, but only by Tajima's *D* and Fay and Wu's *H* for the YRI and by Tajima's *D* for the CEU. These results can readily

be understood in terms of a selective sweep that has proceeded to different stages in the different populations, as will be discussed later.

| Location | Sample characteristics | | | Allele frequency distribution tests | | | | Haplotype test |
|---|---|---|---|---|---|---|---|---|
| | Sample size (chr) | Number of polymorphic sites | Nucleotide Diversity (π) ($\times 10^4$) | Tajima's *D* | Fu & Li's *D* | Fu & Li's *F* | Fay & Wu's *W* | Fu's $F_s$ |
| All populations | 155 | 123 | 4.5 | -2.32* | -2.75* | -3.06** | -46.2** | -27.7** |
| YRI | 52 | 99 | 9.1 | -1.59* | -1.05 | -1.54 | -28.7* | -5.8 |
| CEU | 50 | 7 | 0.5 | -1.57* | -1.17 | -1.54 | -0.9 | -6.6** |
| CHB | 52 | 47 | 1.9 | -2.60** | -3.20** | -3.59** | -33.5** | -5.2 |
| Active (all$^{\alpha}$) | 8 | 61 | 19.7 | | | | | |
| Inactive (all$^{\alpha}$) | 147 | 76 | 2.0 | | | | | |
| Inactive (African) | 46 | 57 | 4.4 | | | | | |
| Inactive (non-African) | 101 | 21 | 0.7 | | | | | |

**Table 12 Summary statistics for CASP12.** $^{\alpha}$All samples (YRI, CEU, CHB and reference sequence). *P<0.05 **P<0.01 (one-sided tests).

Another type of neutrality test examines haplotypes rather than single variable positions. A total of 36 haplotypes were identified, but one haplotype carrying the inactive allele occurred 99 times and accounted for 64% of the sample (and 76% of non-African chromosomes). Fu's $F_s$ test (Fu 1997) shows that significantly fewer haplotypes are found in the whole sample and in CEU than expected under neutrality (Table 12).

We conclude that sequence variation in *CASP12* is significantly different from that expected under neutrality. Departures from neutral expectation at a single locus can arise by demographic processes; however, the neutral model used in these evaluations incorporated the best-fit demographic model. Thus the most likely explanation for all these deviations is positive selection.

#### *4.1.1.4 CASP12 Network*

A median-joining network was constructed to show the relationships between the inferred haplotypes (Figure 33). The eight haplotypes carrying the active allele were all different from one another and from those carrying the inactive allele. All of the inactive allele haplotypes clustered together, with 99 chromosomes at the center of the cluster, 29 one step away, 6 two steps away, and a few more distant. Outside

Africa, the most distant inactive haplotype was only three steps from the center, whereas there was more diversity among the inactive haplotypes in Africa, and not all radiated directly from the central haplotype.

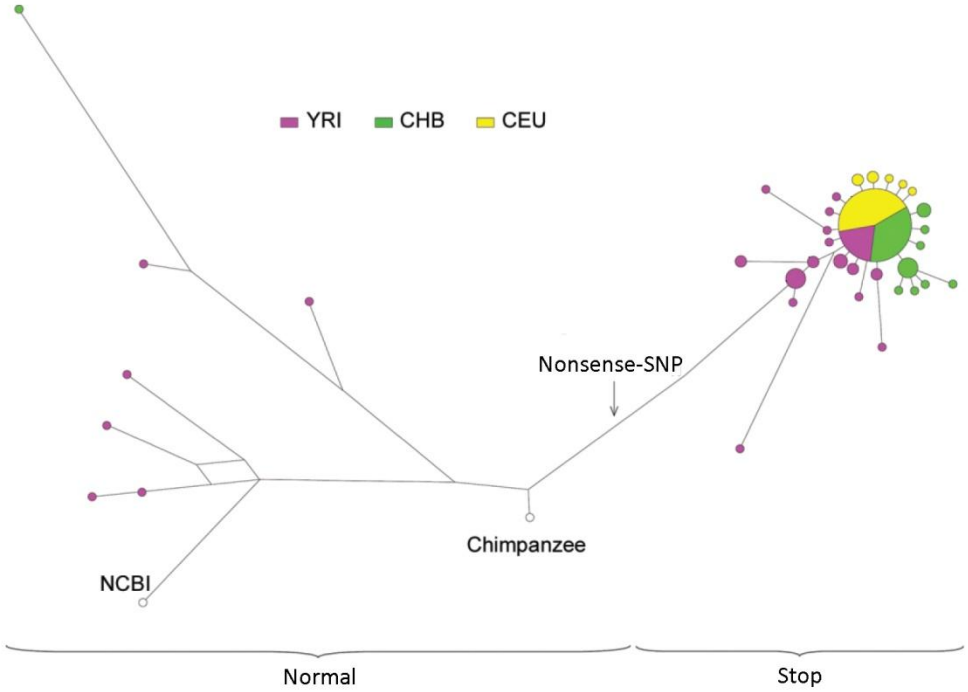


**Figure 33 Median-joining network of inferred *CASP12* haplotypes.** Circle areas are proportional to the haplotype frequency and are colour coded according to population; YRI (purple), CHB (green), and CEU (yellow). Lines represent mutational steps between them; the shortest lines equal one mutation. Location of nonsense-SNP (rs497116) is indicated with an arrow and the cluster of active ("normal") and inactive ("stop") haplotypes is labelled at the bottom. The NCBI reference sequence and the chimpanzee outgroup are labelled.

### 4.1.2 MAGEE2

The *MAGEE2* gene is a melanoma-associated antigen which belongs to a family of *MAGE* genes that are found predominantly on the X chromosome. Several members of the *MAGE* gene family (including *MAGEE2*) are expressed in tumour cells but are silent in normal adult tissues except in the male germ line, leading to an alternative name for these genes, cancer-testis genes. Because of their specific expression on tumour cells, these antigens are potential targets for cancer immunotherapy (Chomez et al. 2001; Ross et al. 2005), but their normal function is completely unknown.

The nonsense-SNP (rs1343879) in *MAGEE2* was identified with the highest $F_{ST}$ value (0.54) in our set of world-wide populations which might be a sign of positive selection. When the $F_{ST}$ value was compared to other empirical $F_{ST}$ values (section 2.3.8.3), it was found to be significantly high (above the 99$^{th}$ percentile) in the HGDP-CEPH populations and in the HapMap. In addition, the geographical distribution of the stop allele (Figure 34) showed an interesting pattern and so taken together this provoked our curiosity about the evolutionary history of the gene. The nonsense-SNP in our dataset was found to truncate the gene by ~77% and yet NMD is not expected to be triggered.

The stop allele (A) has the highest frequency in Asian and Central American populations and is virtually absent from European and African populations. The geographical distribution reveals an east-west gradient of the derived stop allele which may have arisen in the east before the Asian ancestral populations migrated into the Americas less than 20 KYA.

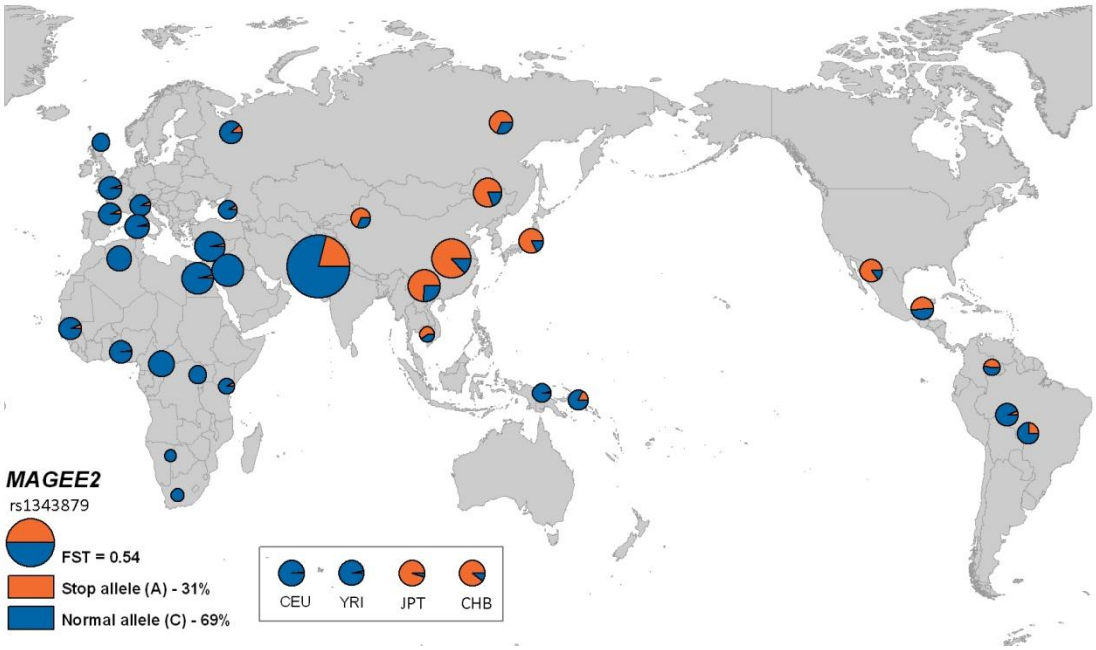


**Figure 34 Geographical distribution of stop (orange) and normal (blue) alleles in *MAGEE2*.** HapMap populations are displayed separately as they do not have precise geographic locations. Pies are proportional to sample sizes.

#### *4.1.2.1 Sequence Variation at MAGEE2*

Next, we wished to explore whether the higher frequency of the *MAGEE2* stop allele in certain populations was the result of population-specific selection or simply of random genetic drift. Therefore, we resequenced a ~12 kb region that covers the whole *MAGEE2* gene and an additional ~5 kb on each side of it in 91 individuals from three HapMap and one extended HapMap population (23 YRI, 23 CHB, 22 CEU and 23 LWK) together with one chimpanzee. 32 chromosomes were found to carry the stop allele – 1 YRI, 28 CHB, 1 CEU and 2 LWK, and these are similar proportions to the worldwide geographical distribution observed in Figure 34. The remaining chromosomes carried the normal allele. A total of 43 SNPs were detected (Appendix F.2.) in the *MAGEE2* gene. We inferred the haplotypes from the SNP data and found that the haplotypes carrying the stop allele were much less diverse than the normal ones, but that the normal were not as diverse as one might have expected (Sachidanandam et al. 2001). Among the 79 chromosomes carrying the normal allele we identified 36 SNPs and a nucleotide diversity ($\pi$) of $3.7\times10^{-4}$,, while the 32 inactive haplotypes only carried 8 SNPs and had a nucleotide diversity of $0.8\times10^{-4}$ (see

summary in Table 13), which was even lower than the diversity of the stop allele chromosomes in *CASP12* $2.0\times10^{-4}$. Again we see a higher diversity in the African populations ($\pi = 4.3\times10^{-4}$ in YRI and $\pi = 4.7\times10^{-4}$ in LWK) compared to the CEU ($\pi = 2.9\times10^{-4}$) and CHB ($\pi = 1.6\times10^{-4}$), but this ratio is not as extreme as that found in the *CASP12* gene (see section 4.1.1.1). The lower diversity observed for the truncated version is consistent with positive selection, but to explore this possibility further we needed to apply more tests.

#### *4.1.2.2 Long-Range Haplotype Test (MAGEE2)*

Unfortunately, the *MAGEE2* nonsense-SNP was not included in our REHH analysis of nonsense-SNPs reported in section 2.3.8.4. This is because *MAGEE2* lies on the X chromosome and the process of phasing haplotypes needs to be done differently to that for autosomal SNPs because of the different copy number of the X in males and females. Therefore, we were unable to use the phased HapMap data for this SNP; male X chromosomes are perfectly phased, but were too few in number. To compensate for these factors we made use of Haplotter (Voight et al. 2006) again to see if this nonsense-SNP was associated with unusually long haplotypes. Again our results are affected by the polymorphic distribution of the SNP in the HapMap populations. The stop allele was virtually absent from CEU and YRI (2.2% and 1.1% respectively) while nearly fixed in the combined CHB+JPT (91%). The results are displayed in Figure 35. The low frequency of the stop allele in YRI made it difficult to make any inferences from the LRH figures which are thus not shown. However, we looked at a region spanning 2.5 Mb around the nonsense SNP in CEU (Figure 35A and B) and CHB+JPT (Figure 35C and D). We found haplotype blocks surrounding the high frequency normal allele in CEU that seems to decay rapidly in about half the chromosomes. In the CHB+JPT, the haplotypes associated with both alleles are long, but decay is noticeably slower for the stop allele and is also much slower than is observed for the normal allele in CEU, which could indicate that recent positive selection has acted on the stop allele in the CHB+JPT populations.

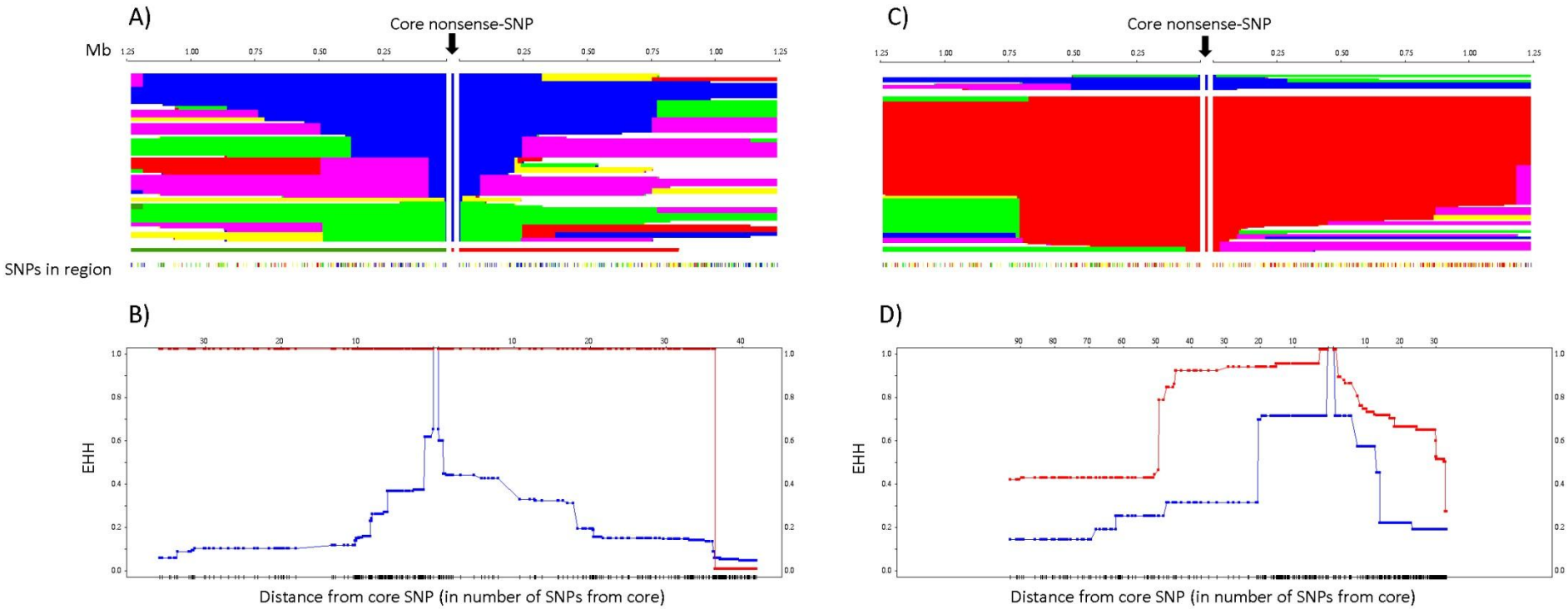


**Figure 35 Long-range haplotypes analysed for region surrounding core nonsense-SNP in *MAGEE2* in CEU and CHB+JPT. Haplotypes at different distances from the nonsense-SNP (core) at the centre are displayed for A) CEU and C) CHB+JPT.** Each horizontal line represents the haplotype of each chromosome. The blue vertical line represents the ancestral state (normal allele) and the derived state (stop allele) is represented in red. The distances over which the haplotypes are spread is displayed at the top of the graph. The total region size displayed is 2.5 Mb and the SNPs in the region are showed at the bottom**. The decay of Extended Haplotype Homozygosity (EHH) at different distances from the nonsense-SNP (core) is displayed for B) CEU and D) CHB+JPT.**

#### 4.1.2.3 Neutrality tests (MAGEE2)

We calculated Tajima's *D* (Tajima 1989c), Fu and Li's *D*, *D**, *F* and *F** (Fu and Li 1993), and Fay and Wu's *H* (Fay and Wu 2000) and neutrality could not be rejected for any of these in either the combined or individual populations, except for Fay and Wu's *H* in the CHB population (see Table 13). Fay and Wu's *H* uses the derived allele frequency spectrum to search for evidence of departures from neutrality (Fay and Wu 2000). This could suggest a population-specific selective sweep acting on the stop allele in the CHB population but not in the others. Therefore, we decided to look at the nucleotide diversity of the inactive haplotype in CHB alone, but found that this was not different from that of the combined populations ($\pi = 1.4\times10^{-4}$).

We then analysed the haplotypes more closely with Fu's $F_s$ test (Fu 1997) and found that significantly fewer haplotypes are found in the whole sample than would be expected under neutrality but not for the individual populations (Table 13).

Overall, most of these tests imply that sequence variation in *MAGEE2* is not significantly different from that expected under neutrality and, except for Fay and Wu's *H*, would be consistent with the idea that the stop allele has risen in frequency in the Asian populations as a consequence of genetic drift.

| | Sample characteristics | | | Allele frequency distribution tests | | | | | | Haplotype test |
|---|---|---|---|---|---|---|---|---|---|---|
| Location | Sample size (chr.) | Number of polymorphic sites | Nucleotide Diversity ($\pi$) ($\times10^4$) | Tajima's D | Fu & Li's D | Fu & Li's D* | Fu & Li's F | Fu & Li's F* | Fay & Wu's H | Fu's Fs |
| All$^{\alpha}$ | 111 | 43 | 4.2 | -1.24 | -2.20 | -2.28 | -2.16 | -2.23 | 0.42 | -27.03** |
| YRI | 26 | 22 | 4.3 | -0.49 | -0.07 | 0.00 | -0.26 | -0.18 | 3.10 | -4.25 |
| LWK | 21 | 21 | 4.7 | -0.24 | 0.26 | 0.29 | 0.12 | 0.16 | 2.93 | -4.15 |
| CEU | 33 | 17 | 2.9 | -0.68 | -1.58 | -1.35 | -1.54 | -1.34 | 1.05 | -2.36 |
| CHB | 31 | 11 | 1.6 | -1.10 | -0.14 | -0.58 | -0.54 | -0.87 | -8.32** | -3.54 |
| Active (all$^{\alpha}$) | 79 | 36 | 3.7 | | | | | | | |
| Inactive (all$^{\alpha}$) | 32 | 8 | 0.9 | | | | | | | |
| Inactive (CHB) | 28 | 7 | 0.8 | | | | | | | |

**Table 13 Summary statistics for MAGEE2.** **P<0.01 (one-sided tests, empirical distribution from the best-fit model). $\alpha$All samples (YRI, LWK, CEU, and CHB).

#### *4.1.2.4 MAGEE2 Network*

A median-joining network was constructed from the inferred haplotypes of *MAGEE2* (Figure 36). As was seen in the geographical distribution of the nonsense-SNP (Figure 34) there is a clear east-west division for the haplotypes which is caused by the nonsense-SNP. All haplotypes carrying the inactive form in the CHB population cluster together (inside red circle in Figure 36) where there is one high-frequency haplotype with the other nonsense-allele haplotypes only one or two steps away. This pattern helps to explain the significantly negative value of Fay and Wu's *H* in the CHB sample by illustrating the moderately high frequency of a derived haplotype cluster specific to the CHB.

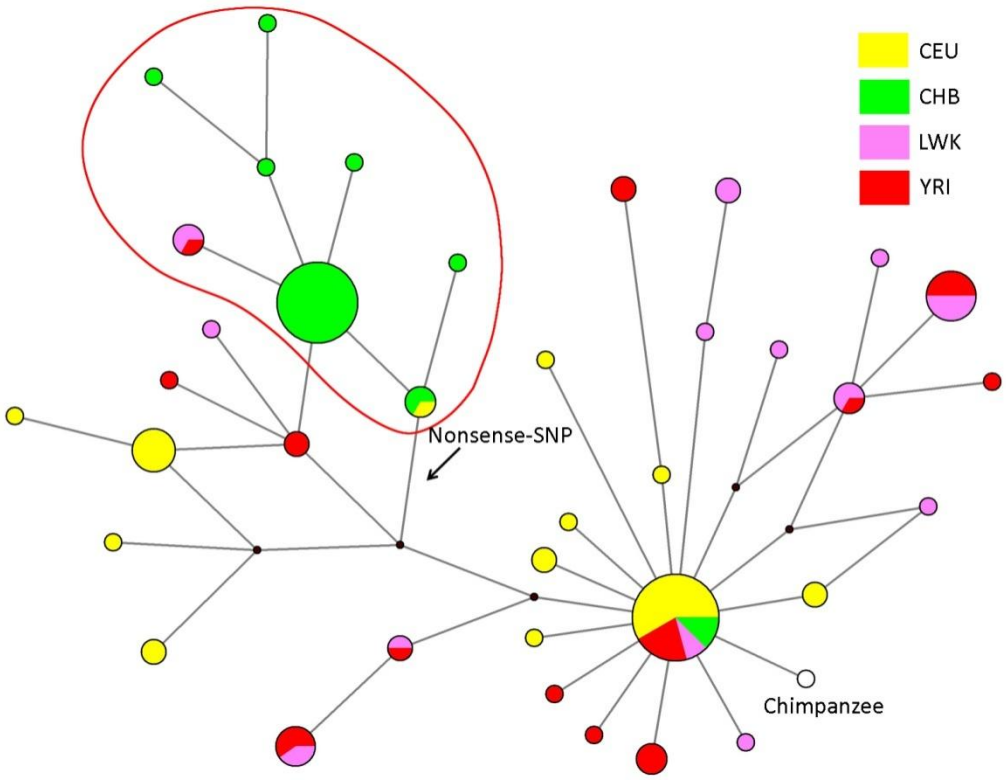


**Figure 36 Median-joining network of inferred *MAGEE2* haplotypes.** Circle areas are proportional to the haplotype frequency and are colour-coded according to population; CEU in yellow, CHB in green, LWK in pink and YRI in red. Lines represent mutational steps between them (one or two steps, according to length). Arrow shows the location of nonsense-SNP (rs1343879).

## 4.2 CONCLUSIONS

These results illustrate the information that can be obtained by more detailed resequencing studies of individual genes. *CASP12* provides one of the clearest examples of positive selection thus far identified in the human genome. Selection has carried the stop allele to a high frequency in the YRI, so significant values of the summary statistics Tajima's *D* and Fay and Wu's *H* are seen. This allele is present at even higher frequency (although not fixed) in the CHB, so shows even more highly significant values of the statistics. In the CEU sample, it is fixed, with the result that diversity is low and the summary statistics have less power, leading to less significant values. The selective sweep is thus proposed to have proceeded to different stages in the different populations. Long Range Haplotype tests show no evidence for selection, a finding that does not conflict with the allele frequency tests but rather is a consequence of the time of selection (too ancient for an LRH signal) and high frequency of the selected allele (too high for an LRH signal). As carriers of the stop allele are protected against severe sepsis, it is reasonable to propose that avoidance of sepsis, and survival if sepsis develops, has been the selective factor.

*MAGEE2* shows a quite different pattern, with only limited evidence for a departure from neutrality and thus for positive selection. While it could be argued that the observations could be explained by drift, it is also worth considering a scenario involving selection. Here, the low frequency of the stop allele in all sequenced samples except the CHB precludes any signal from summary statistics in most populations. In the sequenced CHB, the network shows evidence of rapid expansion of a cluster of haplotypes but the frequency of the stop allele is 31% in the combined populations and the number of SNPs in the sequenced region of the stop allele chromosomes is 8, explaining why the summary statistics are less significant than for the *CASP12* stop allele in the YRI. This low frequency might be due to a more recent origin of the nonsense-SNP in *MAGEE2* than for *CASP12*, and an LRH signal might thus be expected to reveal positive selection. The combined Asian

populations (CHB+JPT) do in fact display differential decay of LD extending over a large region surrounding the nonsense-SNP. There is another test, the XP-EHH (cross population extended haplotype homozygosity) test (Sabeti et al. 2007), which has been designed to detect selective sweeps when the selected allele is fixed in one population but remains polymorphic across other populations. However, in the case of the *MAGEE2*, the gene was included in the study of Sabeti et al. (2007) and listed on the basis of its high $F_{ST}$ value, but was not associated with any unusual long range haplotype signal, cross-population or otherwise (Sabeti et al. 2007 Table S10).

We propose that the stop allele in *MAGEE2* may have undergone recent positive selection in Asian populations, but without any understanding of the normal function of this gene, it is impossible to speculate usefully about the reasons for selection.